Establishment of herbal prescription vector space model based on word co-occurrence

Shouqiang Chen, Yang Chen, Feng Yuan & Xiaowei Chang

The Journal of Supercomputing
An International Journal of High-Performance Computer Design, Analysis, and Use

ISSN 0920-8542
Volume 76
Number 5

J Supercomput (2020) 76:3590-3601
DOI 10.1007/s11227-018-2559-3

Springer

The Journal of Supercomputing (2020) 76:3590–3601
https://doi.org/10.1007/s11227-018-2559-3

Establishment of herbal prescription vector space model based on word co-occurrence

Shouqiang Chen[1] · Yang Chen[2,3] · Feng Yuan[4] · Xiaowei Chang[4]

Published online: 28 August 2018

Abstract

This paper analyzes the establishment of vector space model (VSM) of herbal prescription. Currently, VSM has been frequently used in knowledge acquisition and information retrieval. However, as VSM ignores the association between words, the herbal prescription can only be expressed with single Chinese herb as unit and the potential semantic information in herbal prescription cannot be fully reflected, which limits the clustering result of herbal prescription. This study investigates the significance of word co-occurrence for the research on the formulation theory of herbal prescription, verifies the association between the major function of herbal prescription between word co-occurrence and proposes the word co-occurrence-based VSM and expression method of second-order weighted eigenvalues. It can be concluded that the proposed method can achieve better effect in clustering analysis of herbal prescription comparing traditional VSM-based expression method.

Keywords Vector space model · Word co-occurrence · Association rule · Herbal prescription

1 Introduction

VSM [1], as one of major traditional text representation models, can only express text with single word as unit [2]. It ignores the potential semantic information [3] and thus cannot achieve high accuracy in knowledge acquisition and information retrieval. In

✉ Feng Yuan
yuanfeng623@163.com

1 Center of Hear of the Second Affiliated Hospital of Shandong University of Traditional Chinese Medicine, Jinan, China

2 School of Information Science and Engineering, Shandong Normal University, Jinan, China

3 University of Waterloo, Waterloo, ON, Canada

4 Key Laboratory of TCM Data Cloud Service in Universities of Shandong, Shandong Management University, Jinan, China

recent years, word co-occurrence technique has been frequently adopted to improve information analysis performance.

Word co-occurrence-based analysis mainly refers to calculating the association between words. To improve VSM based on word co-occurrence method, the following two problems should be addressed: (1) how to express the characteristics of word co-occurrence information and (2) how to express the weights of second-order co-occurrence word characteristics. Some scholars improved VSM-based word co-occurrence. Tong et al. [4] proposed the representation model of semantic vector space by investigating the word co-occurrence information in text; Mitrea et al. [5] proposed the representation model of text vector space based on co-occurrence combination using association rules algorithm; Sun et al. [6] calculated the association between words according to similarity of keywords and relevance of co-occurrence words; Liu et al. [7] investigated the word co-occurrence phenomenon and carried out clustering analysis of Web-retrieved results; Hou et al. [8] proposed a semantic information retrieval model based on traditional Chinese medicine ontology by combining the ontology and potential semantic indexing.

This paper mainly focuses on extracting co-occurrence words from herbal prescription and establishing space vector representation method of herbal prescription based on word co-occurrence. In herbal prescription, partial representation form of word co-occurrence is herb pair. Herb pair is a frequently used and stable herb matching form, which is able to explain compound compatibility rule and the scientific connotation within. This paper proposes co-occurrence-based VSM and the representation method of second-order word weighted eigenvalues. Based on this method, the clustering test of the herbal prescription issued by the outpatient doctor from Department of Cardiovascular, Second Affiliated Hospital of Shandong University of Traditional Chinese Medicine was carried out. Results show that the proposed method can achieve better clustering results compared to traditional VSM.

2 Concepts and descriptions

2.1 Vector space model

Salton proposed vector space model and successfully applied it in retrieval system SMART. VSM can convert text into vector, so the text processing issue is converted into vector operation. Using the text frequency dimension reduction of keywords as well as the weights of keywords calculated by *TF-IDF* algorithm, the definition of VSM is shown as [9]: Suppose the corpus D ($D = \{d1, d2, \ldots, dn\}$) consisting of n texts and the set T ($T = \{t1, t2, \ldots, tm\}$) consisting of m keywords are given. The VSM of the jth text d_j is shown as Eq. 1.

$$d_j = \{f_{1j}, f_{2j}, \ldots, f_{mj}\}, \tag{1}$$

where f_{ij} represents the feature weight of keyword t_i in corpus d_j. f_{ij} can be calculated according to Eq. 2.

$$f_{ij} = TF \times IDF \tag{2}$$

$$IDF = \log(n/x) \tag{3}$$

TF represents the word frequency of keyword t_i; *IDF* represents the inverse frequency of keyword t_i, the definition of *IDF* is shown as Eq. 3. In Eq. 3, n represents the total number of corpora and x represents the total number of texts involving keyword t_i.

VSM only uses two statistical distribution measurements (word frequency and inverse frequency) to calculate the weights of keywords, which may lead to the insufficient information in text semantics [10].

2.2 The association between word co-occurrence and subject in herbal prescription

Analyzing from the language structure and characteristics of herbal prescription, word co-occurrence is manifested as herb pair. A herbal prescription consists of multiple herb pairs centered on one or several types of syndromes [11]. Suppose the codex D can be regarded as the corpus ($D = \{d1, d2, \ldots, dn\}$) consisting of n prescriptions and herbal collection T ($T = \{t1, t2, \ldots, tm\}$) consisting of m herbs, then under the syndrome type w, the conditional probability for the occurrence of herb t_i is $p(t_i|w)$, which is defined as below:

Definition 1 Association with the subject. When $p(t_i|w) > \theta$, t_i is associated with subject w; when t_i is associated with the subject w, $p(t_i|w) > \theta$, wherein θ is a threshold value.

Definition 2 Keywords set that is associated with subject w. The definition of keywords set C that is associated with subject w is as follows:

$$C = \{t_i | p(t_i|w) > \theta\} \tag{4}$$

The words that are associated with herbal keywords set C of syndrome type w have the following properties:

Property 1 *If the co-occurrence word (t_i, t_j) is associated with subject syndrome type w, the co-occurrence rate $p(t_i, t_j)$ of such group of co-occurrence words is larger than a certain threshold.*

Proof It is known that $p(t_i|w) > \theta$, $p(t_j|w) > \theta$

$$\begin{aligned} p\left(t_i, t_j\right) &= \sum p(w)p(t_i|w)p(t_j|w) \\ &> \theta^2 p(w) = \sum p(w)p(t_i|w)p(t_j|w) \\ &> \theta^2 p(w) > \delta \end{aligned}$$

Inference 1 If the co-occurrence rate $p(t_i, t_j)$ of co-occurrence words (t_i, t_j) exceeds a certain threshold, then it can be known that such co-occurrence words are the frequently occurred mutual promotion matching, mutual-assistance matching and mutual restraint matching herb pairs with regard to a certain syndrome type.

3 Word co-occurrence-based semantic vector space model (WCSVSM)

3.1 Word co-occurrence weighting

Suppose the herbal prescription corpus D is given, the herb collection T consists of m herb $T = \{t1, t2, \ldots, tm\}$, wherein (t_i, t_j) represents the word co-occurrence constituted by t_i and t_j, which can be recorded by mutual information in Eq. 5:

$$MI(t_i, t_j) = \frac{p(t_i, t_j)^2}{p(t_i) \times p(t_j)}, \tag{5}$$

where $p(t_i)$ and $p(t_j)$ represent the frequency of herb t_i and t_j, respectively; $p(t_i, t_j)$ represents the frequency of word co-occurrence (t_i, t_j); and $MI(t_i, t_j)$ represents the mutual information of word co-occurrence (t_i, t_j).

Word co-occurrence can be calculated by using weighted formula 6, where $weight$ represents the semantic weight of word co-occurrence (t_i, t_j), *TF* represents the frequency of word co-occurrence, *IDF* represents the inverse frequency of word co-occurrence, $MI(t_i, t_j)$ represents the mutual information of word co-occurrence. The definition of formula is shown as below:

$$weight = TF \times IDF \times MI(t_i, t_j) \tag{6}$$

3.2 Design of word co-occurrence extraction algorithm

The herb pair in word co-occurrence was extracted using association rule. The importance of co-occurrence frequency of herb pair (t_i, t_j) was measured by support degree and confidence coefficient, with definition shown as below:

$$support(t_i, t_j) = freq(t_i, t_j) \tag{7}$$

$freq(t_i, t_j)$ represents the co-occurrence frequency of (t_i, t_j) in text set.

$$confidence(t_i, t_j) = MI(t_i, t_j) \tag{8}$$

$MI(t_i, t_j)$ represents the mutual information of word co-occurrence.

The description for word co-occurrence extraction algorithm is shown in Table 1.

Table 1 Extraction algorithm of co-occurrence word

Input: herbal prescription corpus $D = \{d_1, d_2, \ldots, d_n\}$, minimum support degree threshold γ and minimum confidence threshold Ω
Output: Co-occurrence word paired set R
Step 1: Construct transaction set;
Step 2: Scan the entire transaction set; separately calculate $P(t_i)$, $P(t_j)$ and $P(t_i, t_j)$;
Step 3: Calculate $support(t_i, t_j)$ and $confidence(t_i, t_j)$;
Step 4: If the support and confidence are smaller than the set threshold, return directly to step 2. If the support and confidence are greater than the set threshold, add the co-occurrence word pair (t_i, t_j) to the set R and return to step 2. If all the words in the transaction set are traversed, proceed to the next step;
Step 5: Return the co-occurrence word pair set R

Table 2 Building process of WCTVSM

Input: herbal prescription corpus $D = \{d1, d2, \ldots, dn\}$ and herb pair set $R = \{r1, r2, \ldots, rm\}$
Output: Semantic vector space based on co-occurrence word
Step 1: Calculate the mutual information of the herb pairs for all herb pair set;
Step 2: According to Eq. 9, calculate semantic weights for each effective drug;
Step 3: Express the herb prescription corpus into a semantic-based vector space

3.3 Flow design of WCSVSM

Suppose there is a herbal prescription corpus $D = \{d_1, d_2, \ldots, d_n\}$ consisting of n prescriptions. In D, the extracted word co-occurrence set is $R = \{r_1, r_2, \ldots, r_m\}$, where r_m represents the mth extracted word co-occurrence. Then, the herbal prescription corpus can be regarded as a $m \times n$ matrix, where the row vector $d_i = \{dr_{i1}, dr_{i1}, \ldots, dr_{i1}\}$ represents a prescription. In the matrix, the element dr_{ij} represents the distribution of word co-occurrence; if there is word co-occurrence, the weight value is *weight* (the definition of *weight* is shown as Eq. 6); if there is no word co-occurrence, the weight value is 0, as shown by Eq. 9.

$$dr_{ij} = weight \tag{9}$$

By inputting the herbal prescription corpus and herb pair set, the building process of WCSVSM is shown as Table 2.

4 Experiments and analysis

4.1 Experimental data analysis

The data source for this experiment is 5600 prescriptions collected by a doctor from Department of Cardiovascular, Second Affiliated Hospital of Shandong University of Traditional Chinese Medicine, from June 2015 to December 2017. Common syndromes of heart disease include deficiency syndrome and excessive syndrome.

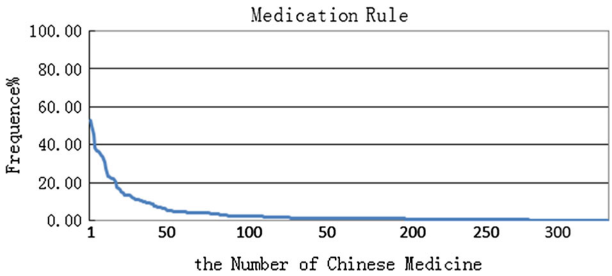

Fig. 1 Frequency chart of Chinese Medicine

Table 3 Part of the frequency table of Chinese Medicine

Chinese Medicine name	Relative frequency (%)	Absolute frequency	Chinese Medicine name	Relative frequency (%)	Absolute frequency
Ligusticum wallichii	53.07	2972	Radices saussureae	50.18	2810
Red-rooted salvia	48.13	2527	Astragalus mongholicus	42.63	2163
Ophiopogon japonicus	40.46	2042	Schisandra chinensis	38.46	2042
Licorice	36.10	2022	Charred triplet	36.10	1921
Forsythia	33.57	1880	Coptis chinensis	30.69	1718
Mother-of-pearl	27.44	1536	Baikal skullcap	23.47	1314
Glycyrrhiza uralensis	22.74	1274	Rehmannia glutinosa	22.38	1253
Earthworm	22.38	1253	Turmeric	20.22	1132
Cuttlebone	17.33	970	Pinellia ternata	16.61	930
Caulis spatholobi	16.25	910	Rhizoma cyperi	15.16	849

Deficiency syndrome includes Yin deficiency, Yang deficiency, qi deficiency, blood deficiency, while excessive syndrome includes blood stasis, yangon, phlegm, which can occur simultaneously. Eight hundred herbal prescriptions were prepared for each of the seven syndromes, constituting a herbal prescription set. Of the total 5600 herbal prescriptions, there are 331 types of herbs, of which the drug-use frequency of Ligusticum wallichii is the highest, which is 2972, while the frequency of mulberry and turmeric is only 1. The diagram of drug-use frequency is shown as Fig. 1, from which we can see 20 herbs with frequency over 15%. The drug-use frequencies of some herbs are shown in Table 3. The general drug-use frequency is shown in Table 4.

4.2 Initialization code

Logistic regression analysis was conducted with these 20 herbs of which the use frequency is large than 15% as independent variables. Through Wald test, likelihood

Table 4 Total frequency table of Chinese Medicine

Interval y (%)	Absolute frequency	Relative frequency (%)	Cumulative relative frequency (%)
$y < 10$	296	89.4	89.4
$10 \leq y < 20$	18	5.5	94.9
$20 \leq y < 30$	7	2.1	97.0
$30 \leq y < 40$	7	2.1	99.1
$40 \leq y < 50$	1	0.3	99.4
$50 \leq y < 60$	2	0.6	100
$60 \leq y < 100$	0	0	100

ratio test and score tests, the p value is always <0.0001, indicating these 20 herbs have statistical significance. In herbal prescription cluster, 20 herbs which have statistical significance and the use frequency $\geq 3\%$ were selected to constitute attribute set. By traversing the herbal prescription set, if the herbal prescription contains such herb, then the attribute value is 1, otherwise 0.

4.3 Extraction results of co-occurrence words

Of co-occurrence words extracted from herbal prescription, there are 38 pairs of co-occurrence words with confidence coefficient over 0.25. Some co-occurrence words of heart disease are shown in Table 5.

Analysis results show that of all co-occurrence words on heart disease, Yin Tonics are more matched with qi-tonifying herb, blood-nourishing medicinal, yin-astringing herb, for example Ligusticum wallichii matching with elecampane [12], elecampane matching with the root of red-rooted salvia [13]; Ligusticum wallichii has the function of replenishing qi to invigorate the spleen [14], elecampane has the function of nourishing Yin and generating body fluid [15] and the root of red-rooted salvia has the function of nourishing the blood and promoting blood circulation [16].

4.4 Index setting for clustering comparative experiment

According to the above steps, the herbal prescription was expressed by VSM and WCSVSM, respectively. Experiments were conducted using k-means clustering algorithm. The clustering results were evaluated using indexes such as clustering precision, recall rate and F measure value. The calculation methods are expressed by Eqs. 10, 11 and 12.

$$\text{Precision } (P) = \frac{\text{The correct amount of clustering}}{\text{The actual amount of clustering}} \tag{10}$$

$$\text{Recall rate } (R) = \frac{\text{The amount of correct clustering}}{\text{The amount of proper clustering}} \tag{11}$$

$$F \text{ measure value } (F) = \frac{2P \times R}{P + R} \tag{12}$$

Table 5 Some co-occurrence words of heart disease

Co-occurrence words	Confidence coefficient	Co-occurrence words	Confidence coefficient
Ligusticum wallichii, radices saussureae	0.48	Ligusticum wallichii, red-rooted salvia	0.48
Radices saussureae, red-rooted salvia	0.43	Ligusticum wallichii, astragalus mongholicus	0.42
Ligusticum wallichii, Ophiopogon japonicus	0.40	Red-rooted salvia, astragalus mongholicus	0.41
Red-rooted salvia, Ophiopogon japonicus	0.39	Ligusticum wallichii, Schisandra chinensis	0.38
Radices saussureae, astragalus mongholicus	0.38	Radices saussureae, Ophiopogon japonicus	0.37
Astragalus mongholicus, Ophiopogon japonicus	0.37	Ophiopogon japonicus, Schisandra chinensis	0.37
Radices saussureae, Schisandra chinensis	0.36	Red-rooted salvia, Schisandra chinensis	0.36
Astragalus mongholicus, Schisandra chinensis	0.36	Ligusticum wallichii, licorice	0.36
Radices saussureae, licorice	0.36	Ligusticum wallichii, charred triplet	0.36
Red-rooted salvia, licorice	0.35	Radices saussureae, charred triplet	0.35
Red-rooted salvia, charred triplet	0.33	Astragalus mongholicus, licorice	0.32
Ligusticum wallichii, forsythia	0.32	Ophiopogon japonicus, licorice	0.32
Ligusticum wallichii, Coptis chinensis	0.30	Radices saussureae, forsythia	0.29
Radices saussureae, Coptis chinensis	0.29	Red-rooted salvia, forsythia	0.29
Red-rooted salvia, Coptis chinensis	0.29	Astragalus mongholicus, charred triplet	0.28
Ophiopogon japonicus, charred triplet	0.28	Schisandra chinensis, licorice	0.28
Licorice, Coptis chinensis	0.27	Schisandra chinensis, charred triplet	0.26
Astragalus mongholicus, forsythia	0.25	Ophiopogon japonicus, forsythia	0.25
Licorice, forsythia	0.25	Charred triplet, forsythia	0.25

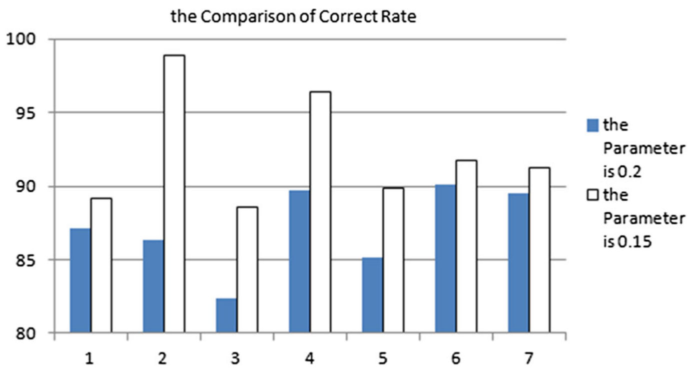

Fig. 2 Comparison diagram of correct rate

4.5 Experiment results analysis

To validate the effectiveness of WCSVSM algorithm improvement, three sets of comparative experiments were designed.

The first set of comparative experiments: to verify the influence of VSM algorithm weight to experimental results. The influence to results was judged according to Eq. 13 (without weights) and Eq. 2 (with weights), respectively.

$$f_{ij} = \begin{cases} 0 & t_j \notin d_i \\ 1 & t_j \in d_i \end{cases} \tag{13}$$

The second set of comparative experiments: to verify the influence of WCSVSM algorithm weight on experimental results. The influence on results was judged according to Eq. 14 (without weights) and Eq. 9 (with weights), respectively.

$$dr_{ij} = \begin{cases} 0 & r_j \notin d_i \\ 1 & r_j \in d_i \end{cases} \tag{14}$$

Table 6 shows the results of the first and second set of comparative experiments, where γ and Ω are both set to 0.15.

As shown in Table 6, WCSVSM weighted algorithm has the highest precision and recall rate in text clustering results.

The third set of comparative experiments: to verify the influence of change of support degree and minimal confidence coefficient threshold to WCSVSM algorithm results. To guarantee the fairness of the experiment, identical k-means clustering algorithm was adopted for both experiments. The contrast figures of precision and recall rate when support degree γ and minimum confidence coefficient threshold Ω are set to 0.25 and 0.15 are shown as Figs. 2 and 3, respectively.

Figures 2 and 3 show the influence of different combinations of γ and Ω to clustering results. It can be concluded that when the support degree γ and minimum confidence coefficient threshold Ω are set to 0.15, the average precision of text clustering is the highest, but the time consumption increases.

Table 6 Experimental results of the first experiment and the second experiment

Pattern of syndrome	VSM (without weights)			VSM (with weights)			WCTVSM (without weights)			WCTVSM (with weights)		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Blood stasis	71.2	81.1	75.8	75.7	83.2	79.3	89.1	88.4	88.7	89.2	92.6	90.9
Deficiency of vital energy	75.3	80.2	77.7	76.1	83.6	79.7	88.6	87.3	87.9	98.9	97.4	98.1
Deficiency of yin	83.1	73.9	78.2	85.3	78.6	81.8	86.6	81.6	84.0	88.6	92.6	90.6
YANG hyperactivity	73.3	83.5	78.1	79.6	86.4	82.9	85.4	88.2	86.8	96.4	91.7	94.0
Deficiency of yang	77.6	80.2	78.9	82.8	85.1	83.9	85.1	89.8	87.4	89.9	92.7	91.3
Deficiency of blood	82.2	76.3	79.1	83.6	79.7	81.6	89.7	84.3	86.9	91.7	95.6	93.6
Phlegm turbidity	80.0	78.7	79.3	87.2	81.5	84.3	90.5	82.8	86.5	91.2	91.8	91.5
Average value	77.5	79.1	78.2	81.5	82.6	81.9	87.9	86.1	86.9	92.3	93.5	92.8

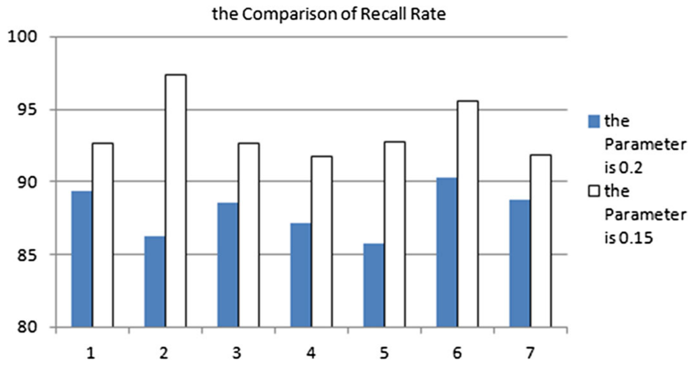

Fig. 3 Comparison diagram of recall ratio

5 Conclusion

This paper analyzes the phenomenon of word co-occurrence in herbal prescription, discusses the association between herbal prescription differential treatment and second-order word co-occurrence and defines the measurement of word co-occurrence weight. The second-order word co-occurrence combination from herbal prescription was extracted using association rules. The semantic space vector model considering word co-occurrence factors was proposed. On the basis text clustering experiments, the comparative experiments with different combinations of support degree and minimum confidence coefficient threshold were designed. Through validation, it can be concluded that the semantic space vector model based on word co-occurrence has stronger semantic expressing capacity and achieves better clustering effect compared with traditional vector space model.

Acknowledgements This paper is financed by the National Social Science Foundation of China (16BGL181) and the Natural Science Foundation of Shandong Province (ZR201702130105).

References

1. Deng Y, Wang Y, Teo KH et al (2015) A simplified space vector modulation scheme for multilevel converters. IEEE Trans Power Electron 31(3):1873–1886
2. Wu FJ, Zhao K, Sun L (2012) Simplified multilevel space vector pulse-width modulation scheme based on two-level space vector pulse-width modulation. Iet Power Electronics 5(5):609–616
3. Jacob B, Baiju MR (2015) A new space vector modulation scheme for multilevel inverters which directly vector quantize the reference space vector. IEEE Trans Ind Electron 62(1):88–95
4. Tong Z, Zhang H (2016) A Text mining research based on LDA topic modelling. In: International Conference on Computer Science, Engineering and Information Technology, pp 201–210
5. Mitrea D, Nedevschi S, Badea R (2018) Automatic recognition of the hepatocellular carcinoma from ultrasound images using complex textural microstructure co-occurrence matrices (CTMCM). In: International Conference on Pattern Recognition Applications and Methods, pp 178–189
6. Sun T, Peng Y, Yang F et al (2017) A self-adaptive text classification method based on multiple word embedding and neural network. Icic Express Lett 11(6):1133–1141
7. Liu AA, Su YT, Nie WZ et al (2017) Hierarchical clustering multi-task learning for joint human action grouping and recognition. IEEE Trans Pattern Anal Mach Intell 39(1):102–114

8. Hou J, Li D, Qiu C et al (2016) A semantic retrieval model based on domain ontology of orchard disease and pests. Chin J Electron 25(3):460–466
9. Saany SIA, Mamat A, Mustapha A et al (2017) Syntax and semantics question analysis using user modelling and relevance feedback. Int J Adv Sci Eng Inf Technol 7(1):329
10. Onan A, Korukoğlu S, Bulut H (2016) A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. Expert Syst Appl 62:1–16
11. Yin W, Schnabel T, Schütze H (2015) Online updating of word representations for part-of-speech tagging. In: Conference on Empirical Methods in Natural Language Processing, pp 1329–1334
12. Giap TH, Dung NA, Thoa HT et al (2018) Phthalides and other metabolites from roots of ligusticum wallichii. Chem Nat Compd 54(1):1–4
13. Stojakowska A, Malarz J, Kiss AK (2016) Hydroxycinnamates from elecampane *(Inula helenium* L.) callus culture. Acta Physiol Plant 38(2):1–5
14. Pu X, Li X, Zhang W et al (2017) Preparation of chitosan/safflower and ligusticum wallichii polysaccharides hydrogel for potential application in drug delivery and tissue engineering. J Mater Res 32(14):1–9
15. Petkova NT, Ivanov I, Vrancheva R et al (2017) Ultrasound and microwave-assisted extraction of elecampane (*Inula helenium*) roots. Nat Prod Commun 12(2):171–174
16. Wang Y, Liang Y, Pediatrics DO (2018) Clinical efficacy of pseudo-ginseng, the root of red-rooted salvia, Hawthorn powder combined with atorvastatin in the treatment of hyperlipidemia. China Med Pharm